Combining Morphological and Ngram Evidence for Monolingual Retrieval

Jaap Kamps Christof Monz Maarten de Rijke

U. of Amsterdam

Our Aims at CLEF 2002

- We took part in the monolingual task for all of the 7 non-English European languages at CLEF
- Our participation was motivated by the following aims:
  - Refine our existing morphological normalization tools, esp. for Dutch and German
  - Develop language independent morphological normalization tools — we experimented with ngrams
  - Experiment with combinations of morphological and ngram-based runs

Experimental Setup

- Homegrown retrieval system called **FlexIR**, using Lnu.ltc weighting scheme and blind feedback
- Morphological runs use lemmatizers where available
  - Dutch: UPLIFT stemmer
  - (Dutch: MBLEM (memory based lemmatizer))
  - French, German, Italian: TreeTagger
  - Spanish: stemmer from Jacques Savoy's site
  - Finnish, Swedish: no tools
- Morph. runs for Du and Ge use decompounding
  - ∼800,000 compounds for Ge, ∼70,000 compounds for Du

Experimental Setup

- Ngram runs
  - ngrams of length $\sim$ avg. word length
  - stopword removal before ngramming
  - ngrams don't cross word boundaries
  - include whole word plus ngrams obtained from it
- Combining runs
  - normalize the retrieval status values (RSVs)
  - interpolate: $\lambda \cdot RSV_1 + (1 - \lambda) \cdot RSV_2$
  - best values of λ determined using the CLEF 2001 test set
  - language/collection dependent

Results

Type	Dutch	French	German	Italian	Spanish
Morphological	0.3673	0.4063	0.4476	0.4285	0.4370
Ngram	0.4542	0.4481	0.4177	0.3672	0.4512
Combined	0.4598	0.4535	0.4802	0.4407	0.4734

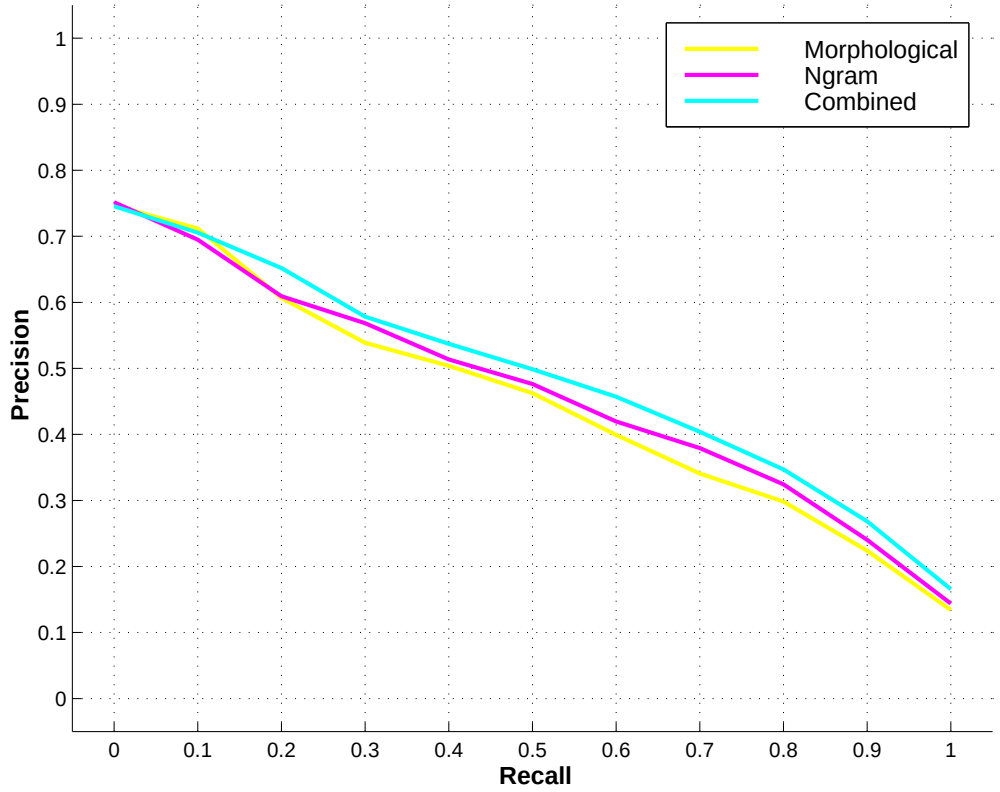
Type	Finnish	Swedish
Morphological	–	–
Ngram	0.3034	0.4187
Combined	–	–

Post-submission Results

Type	Dutch	French	German	Italian	Spanish
Morphological	0.4404	0.4063	0.4476	0.4285	0.4370
Ngram	0.4542	0.4481	0.4177	0.3672	0.4512
Combined	0.4598	0.4535	0.4802	0.4407	0.4734
Combined	0.4760	0.4589	0.4830	0.4422	0.4806
	+4.8%	+2.4%	+7.9%	+3.2%	+6.5%

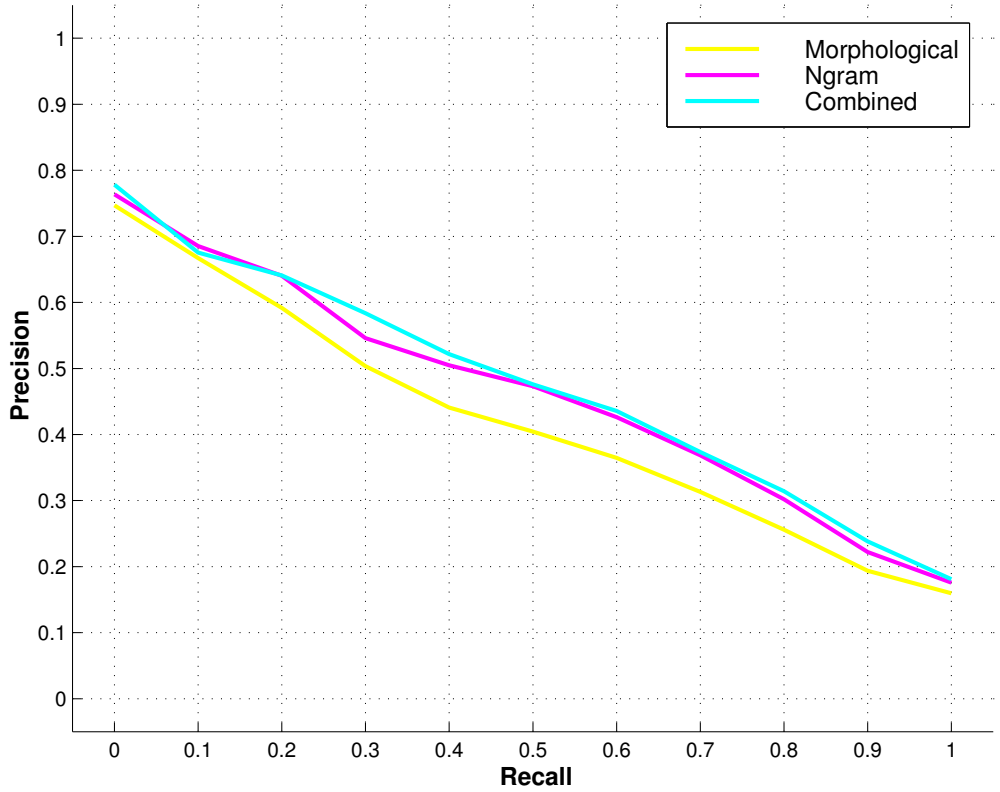
Type	Finnish	Swedish
Morphological	–	–
Ngram	0.3034	0.4187
Combined	–	–

Results: Interpolated avg. precision



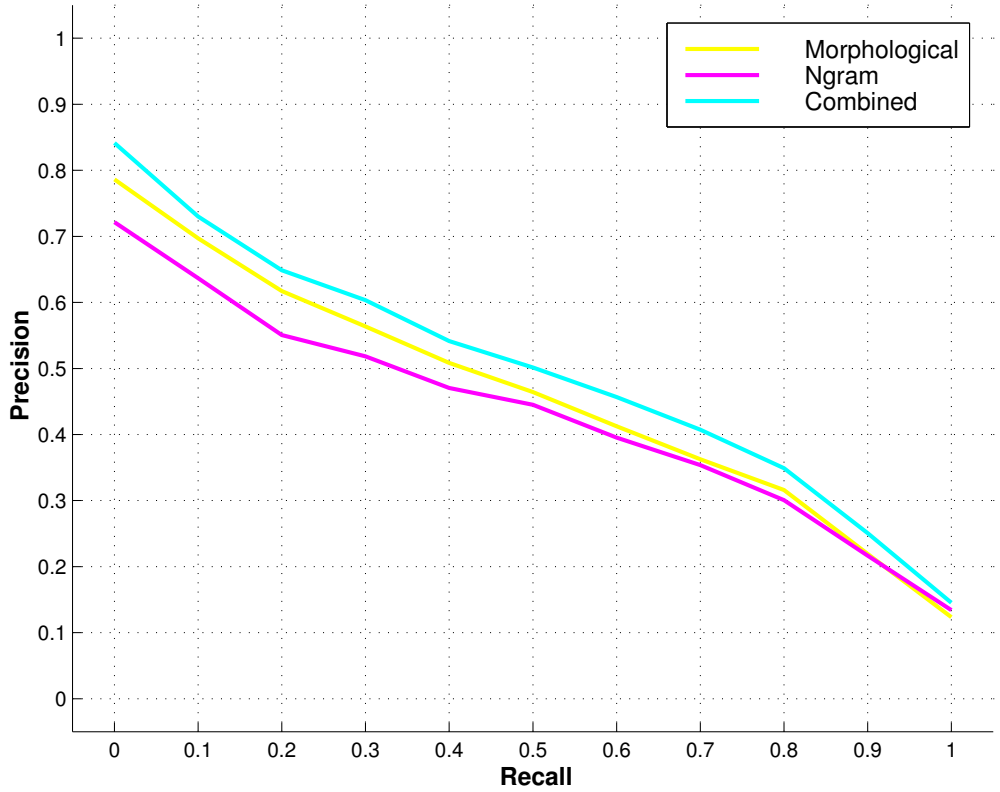
Dutch

Results: Interpolated avg. precision



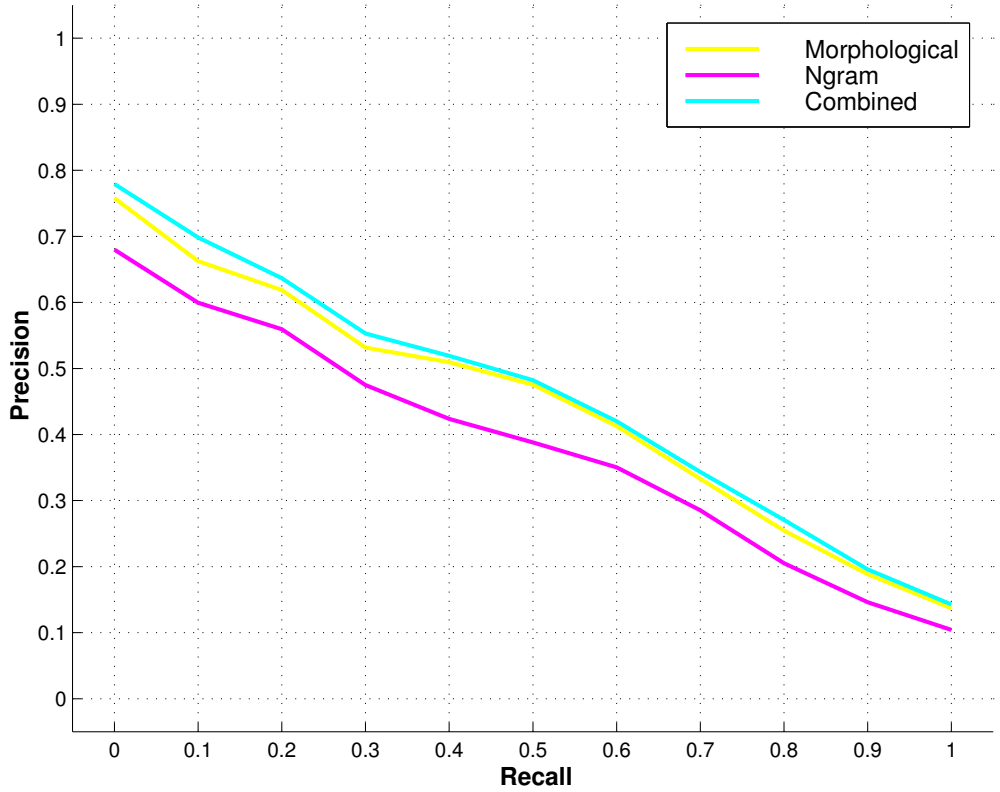
French

Results: Interpolated avg. precision



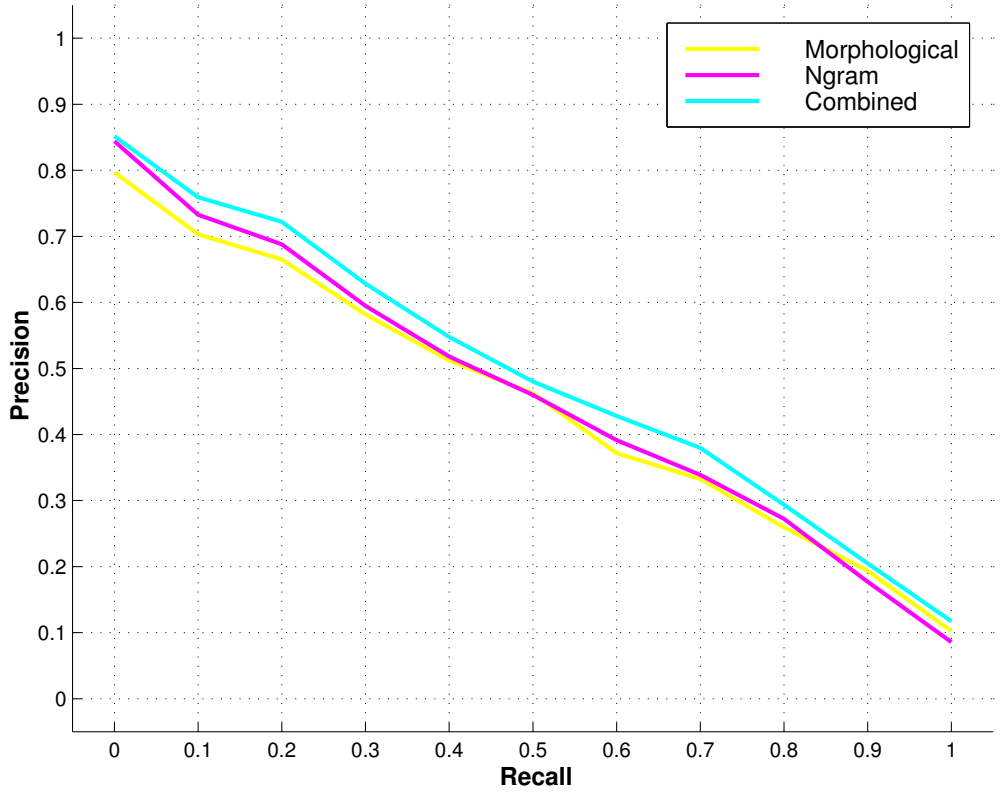
German

Results: Interpolated avg. precision



Italian

Results: Interpolated avg. precision



Spanish

Discussion

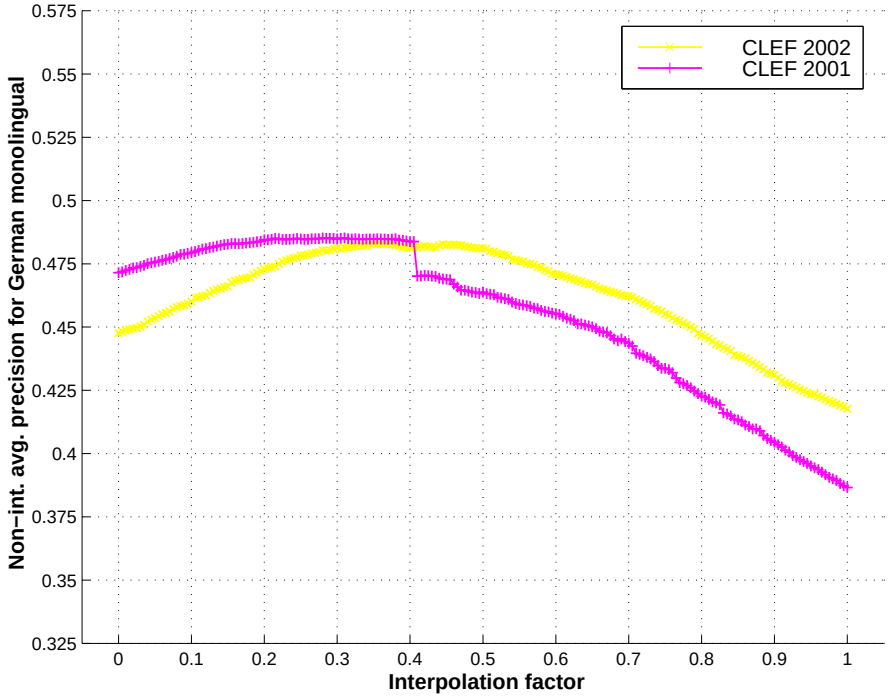
- Rationale for combining runs: maximize overlap of relevant docs between runs, while minimizing overlap of non-relevant docs.
  - Lee's overlap coefficients

$$R_{overlap} = \frac{R_{common} \times 2}{R_1 + R_2} \qquad N_{overlap} = \frac{N_{common} \times 2}{N_1 + N_2},$$

	Dutch	French	German	Italian	Spanish
$R_{overlap}$	0.9443	0.9606	0.9207	0.9021	0.9172
$N_{overlap}$	0.3790	0.5187	0.4180	0.4510	0.5264
Improvement	4.8%	2.4%	7.9%	3.2%	6.5%

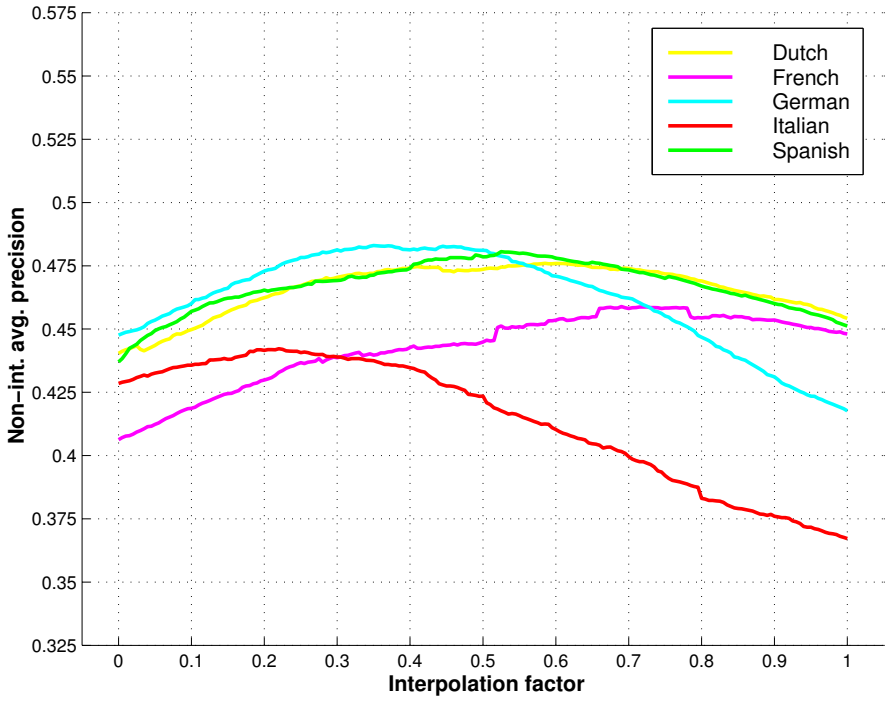
Discussion

- Interpolation factors are robust across topics



Discussion

- Factors can be selected from a broad interval of values without dramatic penalties



Conclusions

- Linguistically informed morphological normalization does improve retrieval effectiveness
- Ngram-based retrieval can be a viable option in the absence of linguistic resources
- Combining runs is a method that can consistently improve base runs

http://www.illc.uva.nl/LIT

http://monzilla.net